PHS 398 XML Schema Guide

Draft Version 1.0
February 20, 2003

Schema Files

The PHS 398 schema is delivered in four files. These are:

commontypes.xsd: Contains the foundation XML vocabulary for the "Research and Related" and "NIH Specific" namespaces (see rarschema.xsd and nihschema.xsd, below).

rarschema.xsd: Contains XML vocabulary to support the "Research and Related" package specification.

nihschema.xsd: Contains the NIH-specific type definitions and reusable element definitions. Some of these definitions are totally separate from the Research/Related elements, and are expressed in a root element named ResearchApplicationExtension. However, many definitions are extensions to the Research/Related elements. The root element that extends the Research/Related "branch" of the document tree is named ResearchAndRelatedApplication. ResearchAndRelatedApplication is comprised of many elements that are taken directly from the R/R schema (rarschema.xsd), but also extends those structures where necessary to accommodate the inflections that are found in the PHS 398 form.

phs398schema.xsd: Contains the overall PHS398 document structure. The root element for the application document is called CompetingGrantApplication, which is comprised of one or more combinations of ResearchProjectDetails and ProjectTypeCode. For the CGAP pilot, there will only be one such combination of elements. The additional cardinality is intended to support future releases, where we will begin to introduce the concept of subproject components, which are submitted together with the primary project application.

One ResearchProjectDetails component will be comprised of a ResearchAndRelatedApplication node as defined in the NIH schema (nihschema.xsd), and a ResearchApplicationExtension, again as defined in the NIH schema.

PHS 398 Crosswalk to XML Document Structure

A supplementary specification has been supplied, in the document entitled ***Mapping of Form PHS 398 to XML Document Structure,*** found in the file named "PHS398Crosswalk.pdf". This document is essentially a very large spreadsheet, with the following columns:

PHS 398 Page: Indicates the PHS 398 form page being mapped. In some cases a data element is expressed in this cross-walk which is not currently collected on paper, but has

been introduced to provide valuable information in the electronic datastream. Such elements are expressed with "PHS 398 Page" indicating "Derived".

PHS 398 Item: Indicates where on the form page the mapped item appears. Labels or other identifying wording from the printed page is used, to the greatest extent possible.

Item Definition: While this is often expressed clearly enough in the name of the item itself, any information that can help clarify what type of information is being collected will be expressed here.

Component Path / Identifier: This indicates where the item should be found or stored, within the document structure defined by the PHS 398 schema (phs398schema.xsd). To be as concise as possible, the CompetingGrantApplication identifier has been implied, for *every* component path shown in this table.

Data Element Name(s): The "Component Path" column takes you to the nearest complex type structure in which the form element will be stored. The "Data Element Name(s)" column indicates the elementary component(s) in which the mapped information will be represented.

Automated Schema Documentation

The schema definitions were all developed with the assistance of the XMLSpy tool. XMLSpy provides a form of automated document generation which is also being supplied with this package. The automated schema document is found in the file named "index.htm". This file will present the XML schema vocabulary and structures in a more graphic layout. The organization of this document is not top-down from the root element. However, knowing the root element (CompetingGrantApplication), the documentation can be easily followed to traverse the tree structure to any level of interest.

Issues for Later Iterations

Several issues turned out to be more complex to deal with in this initial version of the XML schema specification. These will be addressed in a later version. In addressing these issues, some changes may be required to the document structure described herein, although every effort will be made to minimize the impact.

1. A document "core" namespace is still in progress. The "Research And Related" namespace (rarschema.xsd) defines a fair number of the elements that are commonly perceived as the application "core", as represented on standard form SF424. The "Research and Related" namespace identifies these by naming standard, prefixing each "Core" element name with the word "Core". A later draft of this specification will build upon this initial effort and extract the entire Core

into an entirely different namespace, which will then be imported by any schema that needs to build up from the basic Core components.

2. "File Identifiers" or "Attachments" are expressed in several places throughout the mapping document. While this mapping is fairly certain, what is not yet known is the specific nature of such data elements. They may be URN's, or may simply be string identifiers which are repeated as headers on individual file attachments within the same SOAP transaction envelope. This concept will take shape over time and will be expressed in finer detail in later drafts of this documentation.

3. A more stringent application of naming conventions will be applied to these schemas, in later editions. While the names used at this time are certainly descriptive enough, some attempts will be made to apply greater consistency in names across the document structure.

4. The "ResearchBudget" element, as expressed in the "Research and Related" package specification, presented too many difficulties to be properly modeled in this version of the rarschema.xsd. Greater emphasis will be given to this issue in the next phase, to ensure that this element (which appears near the top of the R/R document tree) can be represented properly, and then used as appropriate within the PHS 398 structure.

5. The "Research And Related" package models the individual assurances found on the PHS 398 checklist page as separately-collected data elements, with explanations given for any assurance which cannot be met. In order to expedite the efforts for CGAP pilot phase 1, these have been modeled in the same way as the PHS 398 form and instructions indicate. That is, these assurances are *implied*, in their *entirety*, by the presence of the Signing Official's signature on the application face page. Attempts may be made, in later phases of the XML schema, to model these assurances after the structure represented in the R/R specification.

6. Signatures are currently represented in the XML document structure, as simple strings named "...Authentication". As is the case with file attachment indicators, it is still uncertain at this time what actual form such "authentication" identifiers will take.

7. The document structure modeled here follows the basic "Research Grant" format indicated by the table of contents page (page 3) found in the PHS398 application package. This format should accommodate all of the applications that are expected during the first phases of the CGAP pilot. As the project unfolds to include more complex application types, emphasis will be placed on modeling the other variations of material that may be submitted with such proposals.

8. The structure and mappings in this phase do not make heroic attempts to exceed the scope of the pilot. In other words, the more "exotic" types of applications will

require further analysis and possible XML rework, in later phases. Some types of submissions which the XML structure will *not* support completely at this time are: subprojects, consortiums/contractual arrangements, and SBIR/STTR, to name just a few.